A Federated Dataset of Paraffin Embedded Tissues for caBIG

Michael Feldman, MD, PhD and Rebecca Crowley, MD, MS

9/16/2004

DRAFT: version 1.2

Background:

All caBIG participating sites sit on a mountain of textual reports and tissue in the form of pathology reports and paraffin blocks. Access to these archives has enormous potential to facilitate research at individual institutions as well as across the participating sites within the caBIG community and the larger cancer research community. The use of institutional paraffin archives provides an immediate source of tissue information and resources that could be leveraged for purposes within the TBPT workspace, but also for the other two domain workspaces.

To provide access to this information, we are proposing an extension of the Shared Pathology Informatics System (SPIN) to the caBIG community. Similar to SPIN, shared information would be de-identified (stripped of the 18 HIIPA identifiers). As in the SPIN Network, local repositories will maintain identifiers, which provide access to the tissue in their own paraffin archives. However, these identifiers will not be available to other institutions. Extending work performed n the SPIN project, the developing caTIES application will utilize NCI resources, such as the caDSR, NCI Thesaurus, and NCI Metathesaurus to perform the coding and retrieval of documents and data (see description of caTIES below).

To maximize our chances for success – we propose a multiphase plan which will move from local repositories towards federated access over a period of years. We propose to utilize the developing caTIES system to create a process that would allow:

#1 Local repositories with search tool that only accesses internal data – Year 1

#2 Local repositories with search tool that queries and returns aggregate data across the grid – Start in year 1, Finish year 2

#3 Local repositories with search tool that queries and returns individual reports. Are these deidentified enough? Start year 2

#4 Individuals at other institutions requesting tissue - Start year 3

Access to the search tool, in and of itself does not provide access to the tissues. This, in our minds is a separate issue and will need to be dealt with on an individual case basis. This federated archive is seen as a way for researchers to mine information, as well as a resource to find potential cases for research. The retrieval and sharing of any tissues identified by this search tool, would then need to vetted through individual institution IRB's as well as material transfer agreements.

As can be seen from the above goals, these are achievable at different level of complexity and would be addressed at different years of the project. The most difficult (#3 and #4) require a good deal of thought as to how to properly de-identify information, how to identify patients across institutions and how to share materials. This proposal does not have the answers to these questions, however, as a group we must work to deal with these issues whether related to the caTIES project or to caTissue as they will be central issues to either project.

This project raises many questions that will need to be dealt with by the group, including but not limited to the following:

a. Multiple levels of access to data – defined by IRB protocols and need for information. Must be subsumed within the software design? Local access vs grid access?
b. How is access granted, who decides on granting the privileges?
c. What data can be retrieved? De-identified report? Tissue blocks? Slide recuts?
d. IRB issues – patient safety/confidentiality – Can we define a single standard by which all IRB's will approve? Can we create a templatable approach for investigators to use to submit to IRB for approval?
e. Material transfer – not an issue until we tackle task #4 but a major impediment to sharing resources. Handle on a global scale or deal with it on a case-by-case collaborative project?
f. How are material transfers handled? What is process? Who pays?
g. How do we build the application to satisfy all IRB's without the need to customize for every IRB?

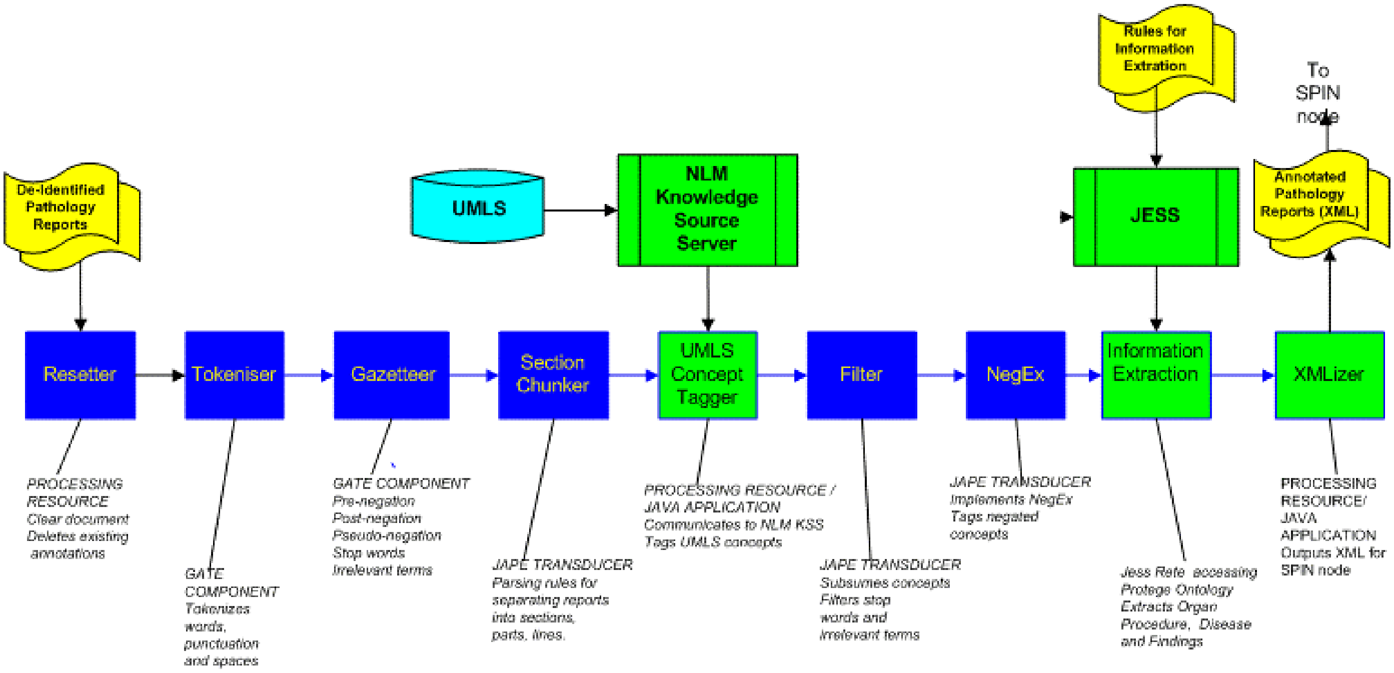
Access to information through free-text reports

Access to cases (both the information in the surgical pathology report and the tissue itself) requires the ability to search and retrieve information from the distributed archive on the basis of search terms derived from a controlled vocabulary and relationships between those concepts, specified in a repository of common data elements.

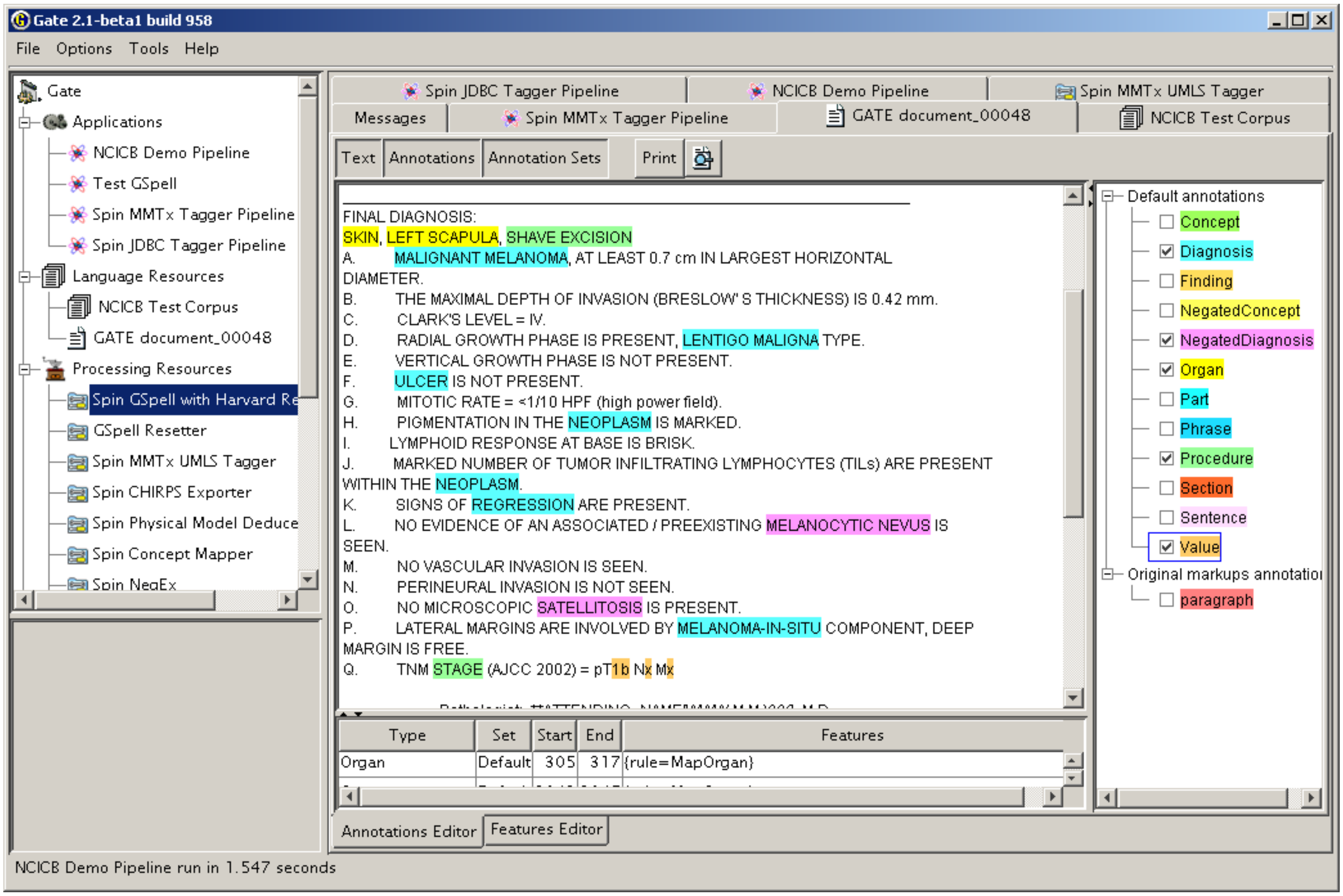
For the Distributed Archive, we propose to utilize caTIES - a development project that has been proposed by University of Pittsburgh jointly for the Vocabularies and CDE and TBPT workspaces. In phase 1 of the project, CaTIES will provide the (1) the text-processing capabilities, (2) the database to enable storage and retrieval of annotated documents using NCI Metathesaurus terms, as well as selected key features of these documents such as site, diagnosis, and TNM stage, (3) caTIES will provide an initial query interface for searching and aggregating data from the archive. In phase 2, caTIES will provide a more sophisticated query system that also permits distributed queries implemented across an OGSA-DAI – based data grid.

CaTIES is an extension of work done for the Shared Pathology Informatics System – and NCI funded collaboration to develop tools for accessing tissue and information from surgical pathology cases across institutions. The system is built with GATE - an open source Java framework for language engineering – (http://gate.ac.uk/).

The pipeline accomplishes the automated sequential processing of surgical pathology reports from free-text to coded data using the following modules. For caTIES we will replace the UMLS Metathesaurus with the NCI Metathesaurus.


Rules for Information Extration
To SPIN node
De-Identified Pathology Reports
UMLS
NLM Knowledge Source Server
JESS
Annotated Pathology Reports (XML)
Resetter
Tokeniser
Gazetteer
Section Chunker
UMLS Concept Tagger
Filter
NegEx
Information Extraction
XMLizer
PROCESSING RESOURCE
Clear document
Deletes existing annotations
GATE COMPONENT
Tokenizes words, punctuation and spaces
GATE COMPONENT
Pre-negation
Post-negation
Pseudo-negation
Stop words
Irrelevant terms
JAPE TRANSDUCER
Parsing rules for separating reports into sections, parts, lines.
PROCESSING RESOURCE / JAVA APPLICATION
Communicates to NLM KSS
Tags UMLS concepts
JAPE TRANSDUCER
Subsumes concepts
Filters stop words and irrelevant terms
JAPE TRANSDUCER
Implements NegEx
Tags negated concepts
Jess Rete accessing Protege Ontology
Extracts Organ Procedure, Disease and Findings
PROCESSING RESOURCE/ JAVA APPLICATION
Outputs XML for SPIN node

An example report in a case of Melanoma is shown below in GATE.



FINAL DIAGNOSIS:
SKIN, LEFT SCAPULA, SHAVE EXCISION
A. MALIGNANT MELANOMA, AT LEAST 0.7 cm IN LARGEST HORIZONTAL DIAMETER.
B. THE MAXIMAL DEPTH OF INVASION (BRESLOW' S THICKNESS) IS 0.42 mm.
C. CLARK'S LEVEL = IV.
D. RADIAL GROWTH PHASE IS PRESENT, LENTIGO MALIGNA TYPE.
E. VERTICAL GROWTH PHASE IS NOT PRESENT.
F. ULCER IS NOT PRESENT.
G. MITOTIC RATE = <1/10 HPF (high power field).
H. PIGMENTATION IN THE NEOPLASM IS MARKED.
I. LYMPHOID RESPONSE AT BASE IS BRISK.
J. MARKED NUMBER OF TUMOR INFILTRATING LYMPHOCYTES (TILs) ARE PRESENT WITHIN THE NEOPLASM.
K. SIGNS OF REGRESSION ARE PRESENT.
L. NO EVIDENCE OF AN ASSOCIATED / PREEXISTING MELANOCYTIC NEVUS IS SEEN.
M. NO VASCULAR INVASION IS SEEN.
N. PERINEURAL INVASION IS NOT SEEN.
O. NO MICROSCOPIC SATELLITOSIS IS PRESENT.
P. LATERAL MARGINS ARE INVOLVED BY MELANOMA-IN-SITU COMPONENT, DEEP MARGIN IS FREE.
Q. TNM STAGE (AJCC 2002) = pT1b Nx Mx

Type	Set	Start	End	Features
Organ	Default	305	317	{rule=MapOrgan}

Default annotations
- □ Concept
- ☑ Diagnosis
- □ Finding
- □ NegatedConcept
- ☑ NegatedDiagnosis
- ☑ Organ
- □ Part
- □ Phrase
- ☑ Procedure
- □ Section
- □ Sentence
- ☑ Value

Original markups annotatio
- □ paragraph

NCICB Demo Pipeline run in 1.547 seconds

The data produced from reports results in XML, as shown below for a case of Prostatic Adenocarcinoma. The XML that will be emitted from caTIES is likely to look somewhat different because in addition to data it will carry caBIG metadata. Ultimately data will populate caBIG data structures, to be retrieved by multiples applications across the grid.

```
 <Diagnosis>
   <Term Assertion="Affirmation" Code="C0007112" Source="UMLS">PROSTATIC
       ADENOCARCINOMA</Term>
   <Modifier Assertion="Affirmation" Code="C0205281" Source="UMLS">INVASIVE</Modifier>
   <Modifier Assertion="Affirmation" Code="C0205616" Source="UMLS">MODERATELY
       DIFFERENTIATED</Modifier>
   <Modifier Assertion="Affirmation" Code="C0332207" Source="UMLS">ACINAR</Modifier>
   <Modifier Assertion="Affirmation" Code="C0332307" Source="UMLS">TYPE</Modifier>
   <Modifier Assertion="Affirmation" Code="C0205195" Source="UMLS">COMBINED</Modifier>
   <Modifier Assertion="Affirmation" Code="C0332326" Source="UMLS">GLEASON SCORE</Modifier>
- <Topology>
   <Term Assertion="Affirmation" Code="C0033572" Source="UMLS">PROSTATE</Term>
     </Topology>
  </Diagnosis>
- <Diagnosis>
   <Term Assertion="Affirmation" Code="C0332326" Source="UMLS">GLEASON SCORE OF
       3+3=6</Term>
   <Modifier Assertion="Affirmation" Code="C0332326" Source="UMLS">3</Modifier>
   <Modifier Assertion="Affirmation" Code="C0332326" Source="UMLS">3</Modifier>
   <Modifier Assertion="Affirmation" Code="C0332326" Source="UMLS">6</Modifier>
     </Diagnosis>
```

Query Interface

The caTIES proposal specifies two development cycles, during which the query system will mature from simple to more complex queries. The query interface to the system will initially provide by Metathesaurus term and free-text, and will eventually permit users to use concepts and relationships drawn directly from caDSR. In Phase I we will support simple queries on diagnosis, location, procedure, and finding, and will allow users to specify whether term is explicitly negated. In Phase II, we will permit more complex queries based on relationships within the caDSR or NCI Thesaurus. Examples of queries for these two phases are shown below.

Example Phase I query:

Return all reports where (AGE = 0-18) and (DIAGNOSIS= CL024452: Astrocytic Tumor) and (PROCEDURE= C0740294: Biopsy brain) and (NEGATED-FINDING= C0027540: Necrosis).

Example Phase II query:

Return all reports where (AGE = 0-18) and (DIAGNOSIS= CL024452: Astrocytic Tumor OR children of that NCI Thesaurus concept) and (PROCEDURE= C0740294: Biopsy brain) and (TNM = T2)